

Strategy for investments from Zipf law(s)

M. Ausloos¹ and Ph. Bronlet

SUPRAS and GRASP, B5, Sart Tilman Campus, B-4000 Liège, Euroland

Abstract

We have applied the Zipf method to extract the ζ' exponent for seven financial indices (DAX, FTSE; DJIA, NASDAQ, S&P500; Hang-Seng and Nikkei 225), after having translated the signals into a text based on two letters. We follow considerations based on the signal Hurst exponent and the notion of a time dependent Zipf law and exponent in order to implement two simple investment strategies for such indices. We show the time dependence of the returns.

PACS numbers: 05.45.Tp, 89.65.Gh, 89.69.+x

1 Introduction

Usually analysts recommend investment strategies based e.g. on "moving averages", "momentum indicators", and the like techniques.[1, 2] As soon as econophysicists discovered scaling laws in financial data, it was of interest to search for some predictive value from the laws through some extrapolated evolution. E.g. a technique known as *detrended fluctuation analysis* (DFA) which measures the deviation of correlated fluctuations from a trend was developed into a strategy known as the *local* (or better *instantaneous*) DFA in order to predict fluctuations in the exchange rates of various currencies, Gold price and other financial indices. [5, 6] The statistical analysis of data was

¹e-mail address: marcel.ausloos@ulg.ac.be

based on the value of the exponent of the so found scaling law, itself related to the fractal dimension of the signal, or also to the Hurst exponent of the so called rescaled range analysis. Mathematical extensions, so called q – order DFA and multifractals, can be found in the literature, though optimization problems and predictions on the future of fluctuations are apparently not so evident from these methods. A drawback in the DFA is found in the fact that it rather looks at correlations in the sign of fluctuations rather than at correlations in their amplitude.

Another sort of data analysis technique is known as the Zipf technique,[3, 4] originating in work exploring the statistical nature of languages. The Zipf analysis technique has also been used outside linguistic, financial and economic fields.[7] The technique is based on a *Zipf plot* which expresses the relationship between the frequency of words (more generally, events) and the rank order of such words (or events) on a log-log diagram; a *cumulative* histogram can be drawn as well. The slope of the best linear fit on such a plot corresponds to an exponent s describing the frequency P of the (*cumulative*) occurrence of the words (or events) according to their rank R through, e.g. $P(> R) \simeq R^{-s}$.

There are many instances in which financial and other economic data can be described through a log-log (Zipf) plot : e.g., the distribution of income (Pareto distribution) [8], the size of companies [9], sociology [10], sometimes after translating the financial data into a text [11, 12, 13, 14, 15]. Thus it seems of interest to check whether such a technique can have some predictive value in finance. The present report is in line with such previous investigations. We present results based on considerations that financial data series are similar to fractional Brownian motion-like time series, and usually biased.[16] We examine whether a time dependent Zipf law and exponent exist and can be used in order to implement simple investment strategies.

First, it is thus necessary to translate the financial data into a *text*, based on an alphabet with k characters and search for words of length m . There are obviously k^m possible words. They can be ranked according to their frequency on a log(frequency)-log(rank) diagram. A linear fit leads to consider the relationship as a power law. Moreover, in the spirit of the local DFA, a local (or "time" dependent) Zipf law or exponent can be introduced.[16] In this latter reference, we have also considered the effect of a linear trend on the value of the Zipf exponent.

Here below we have translated seven financial index signals (DAX, FTSE; DJIA, NASDAQ, S&P500; Hang-Seng and Nikkei 225) each into a text based

Table 1: Typical financial indices characteristics between Jan. 01, 1997 and Dec. 31, 2001: Hurst exponent H , (5,2)-Zipf exponent, bias (ε), p_u , p_d , linear trend slope

	H	$\zeta_{(5,2)}$	ε	p_u	p_d	Trend
1. DAX	0.51±0.01	0.11±0.05	0.0332	0.5332	0.4668	2.24±0.07
2. DJIA	0.46±0.04	0.11±0.02	0.0172	0.5172	0.4828	2.97±0.07
3. FTSE	0.43±0.03	0.19±0.08	0.0149	0.5149	0.4851	0.90±0.05
4. Hang-Seng	0.47±0.02	0.08±0.02	0.0060	0.5060	0.4940	1.75±0.20
5. Nasdaq	0.56±0.03	0.19±0.08	0.0428	0.5428	0.4572	1.19±0.06
6. Nikkei225	0.47±0.06	0.15±0.04	-0.0204	0.4796	0.5204	-4.69±0.16
7. S&P	0.51±0.04	0.12±0.03	0.0152	0.5152	0.4848	0.38±0.01

on two letters u and d . Based on the above considerations we have imagined two simple investment strategies, and report on the results (or "returns"). From the beginning we stress that a restriction to two letters is equivalent to examine only correlations in the fluctuation signs. However the Zipf method main interest is surely the capability to consider amplitude fluctuations, - by defining various fluctuation ranges.

2 Data analysis

The daily closing values of (DAX, FTSE; DJIA, NASDAQ, S&P500; Hang-Seng and Nikkei 225) indices, from Jan. 01, 1997 till Dec. 31, 2001 (Fig.1) have been obtained from <http://finance.yahoo.com/>. They contain *ca.*1250 data points. After translating the financial time series into a text, one searches for words, and rank them according to their frequency. On a log-log paper, the best line fit slope is the Zipf exponent. Elsewhere we have already shown that the usual Zipf exponent ζ [3, 4] depends on the normalization process used to calculate the ranks. If the frequency f of occurrence is normalized with respect to the theoretical one f' , i.e. that expected for pure (stochastic) Brownian processes, one has $f/f' \sim R^{-\zeta'}$. The theoretical frequency expected for a letter in a text based on a binary alphabet, u, d takes into account the number n of characters, say of type d (and u), in a word. Suppose that in the text, the frequency of a d (u) letter is p_d (p_u). Usually, a bias exists, i.e. $p_u \neq p_d$. Therefore $f' = p_u^{(m-n)} p_d^{(n)}$. Whether or not the ζ

and ζ' exponent depend on the bias has been examined elsewhere.[16] The p_u and p_d values for the seven indices are reported in Table 1, together with the bias defined here as $\epsilon = p_u - 0.5$. The linear tendency for the time interval is also given in Table 1. We have calculated overall Zipf exponent values, $\zeta_{(m,k)}$, and give the $\zeta_{(5,2)}$ value for the seven indices in Table 1.

In the spirit of the so called local (or better *instantaneous*) DFA method. We can consider that a Zipf exponent is time dependent, thus obtain a *local* Zipf law and *local* Zipf exponent. Only the case for $m < 8$ letter words has been considered, but are not shown for lack of space. This m value is so chosen within the financial idea background having motivated this study, e.g. $m = 5$ is the number of days in a (bank) week !

In general a (one dimensional) financial index can be characterized by a so called Hurst exponent H , obtained as follows. The time series is divided into boxes of equal size, each containing a variable number of "elements". The local fluctuation at a point in one box is calculated as the deviation from the mean in that box. The *cumulative* departure up to the j^{th} -point in the box is next calculated in all boxes. The rescaled range function is next calculated from the difference between the *maximum* and the *minimum*, i.e. the range in units of the rms deviation in the box. The average of the rescaled range in all boxes with an equal size n is next obtained and denoted by $\langle R/S \rangle$. The above computation is then repeated for different box sizes s to provide a relationship between $\langle R/S \rangle$ and s , - which is expected to be a power law $\langle R/S \rangle \simeq s^H$ if some scaling range and law exist.

If $H = 1/2$ one has the usual Brownian motion. The signal is said to be persistent for $H > 1/2$, and antipersistent otherwise. We have calculated the Hurst exponent [17] by this rescaled range analysis [18] for the seven financial index signals. Their H value and the corresponding error bar are given in Table 1. The error bars are those resulting from a best linear fit and a root mean square analysis.

Tests (not shown here) of the stochasticity (or not) of the data can be based on the surrogate data method [19] in which one randomizes either the sign of the fluctuations or shuffles their amplitude, and finally observes whether the error bars (or confidence intervals) of the raw signal and the surrogate data signal overlap.

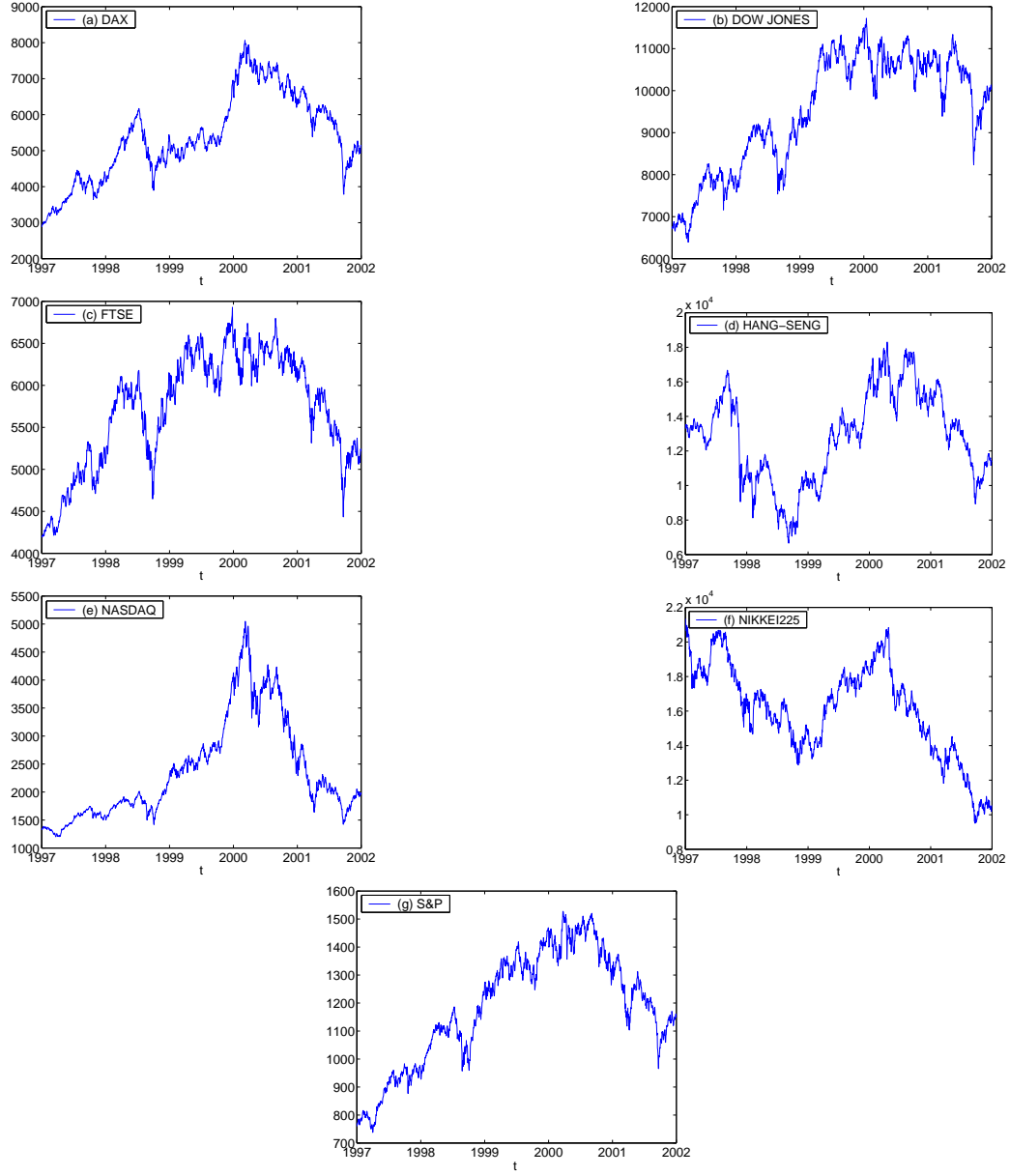


Figure 1: The (DAX, FTSE; DJIA, NASDAQ, S&P500; Hang-Seng and Nikkei 225) indices have been obtained from [http : //finance.yahoo.com/](http://finance.yahoo.com/). They contain *ca.*1250 data points, from Jan. 01, 1997 till Dec. 31, 2001

3 Returns and basic Zipf strategy

The method is based on searching for the probability of a character sequence at the end of a word. We consider the case of what can happen the *next day* after a few ($m - 1$) days) only. Consider a word of length $m - 1$, and calculate in all boxes of size τ the probabilities $p_u(t)$ and $p_d(t)$ to have a character sequence $(c_{t-m-3}, \dots, c_{t-1}, u)$ and $(c_{t-m-3}, \dots, c_{t-1}, d)$ respectively, where c_t represents the character at time t . Since only a $k=2$ alphabet is used, it is fair to develop a simple strategy based *only* on the sign of the fluctuations, thus use a strategy similar to that one implemented in the "instantaneous" DFA, i.e. when expecting correlated or anticorrelated fluctuations, in u and d . In order to avoid investment activity when the choice probability is low we have used a *strength* parameter for measuring the relative probabilities, i.e.

$$K(t) = \left| \frac{p_u(t) - p_d(t)}{p_u(t) + p_d(t)} \right|, \quad (1)$$

varying between 0 and 1, its value giving the number of shares bought (or sold) at a certain investment time.

Table 2: Final returns $r(t)$ in (%) obtained after 5 years on various indices from various strategies Z1 and Z2 as described in the text when based on a $(m, k=2)$ Zipf exponent as compared (second column) to the mere final index value change

$\tau = 500$	$r(t)$	Zipf1			Zipf2		
		$m = 3$	$m = 5$	$m = 7$	$m = 3$	$m = 5$	$m = 7$
1. DAX	77.55	61.33	71.34	62.39	65.01	65.42	32.78
2. DJIA	49.49	47.42	55.73	29.12	34.52	39.12	23.68
3. FTSE	24.30	33.34	30.63	42.98	18.99	29.33	40.92
4. Hang-Seng	-15.48	-23.88	-26.61	-25.13	3.78	-9.81	-9.30
5. Nasdaq	46.55	56.30	39.87	61.41	56.52	65.07	43.85
6. Nikkei225	-50.61	-13.78	-12.39	-8.73	-5.47	-14.60	-7.89
7. S&P	51.16	67.80	80.34	107.79	43.65	57.85	71.33

Results are reported when windows (boxes) of size $\tau = 500$ respectively are moved along the signal. This value corresponds to a 2 year type investment

window. Notice that the *local* exponents are usually larger than the average one, due to finite size effects.

In the *Zipf*₁ (*Z1*) strategy, we consider that if $p_u(t) > p_d(t)$, a "buy order" is given. A "sell order" is given for $p_u(t) < p_d(t)$. No order is given when both probabilities are equal. Results reported in Table 2 pertain to $m=3, 5$, and 7 at the end of the 5 year interval. In the *Zipf*₂ (*Z2*) strategy the local linear trend is subtracted before calculating $p_u(t)$ and $p_d(t)$. The time dependent returns for *Z1* and *Z2* in the case $m=3, 5$, and 7 , and for $k=2$ are given in Fig. 2 for the seven hereby considered financial indices. A return $r(t)$ (given in %) is defined from

$$Bq(t) = Bq(t_0) [1 + r(t)], \quad (2)$$

where $Bq(t)$ and $Bq(t_0)$ are the amount of money available at time t and at the beginning t_0 of the investment period respectively, for a share of value $q(t)$ bought $q(t_0)$ at the starting date.

4 Conclusions

It appears (Table 2) that there is no immediate simple and general rule or universal optimum strategy. The latter depends on the volatility, i.e. the signal roughness and the local (m, k) -Zipf exponent value. From the implemented simple strategies, it occurs that "the best returns" are usually for *Z1* with $m=5$, except for NASDAQ for which a fine result arises from a *Z1* with $m=7$, (or *Z2* and $m=5$) and for the FTSE, with either *Z1* or *Z2* and $m=7$. This choice of the m value and the *Z1* strategy is conjectured to be good for large ζ' and "non Brownian" (large H) cases. However for quasi-Brownian signals (and high/low ζ'), then it is obvious that one has reduced losses for the NIKKEI when one chooses a *Z1* strategy with $m=7$; this is a very large ζ' case. This is rather similar to the FTSE case. Increased gains are found for *S&P* with *Z1* and $m=7$, and for DJIA with *Z1* and $m=5$, i.e. when ζ' is close to 0.1. On the contrary for the *HS* a *Z2* and $m=3$ strategy should be better, i.e. for $\zeta' \ll 0.1$. The situation is rather neutral for the DAX, the choice *Z1*, $m=5$ being favored.

Many other cases could be further considered, and theoretical work suggested : first one could wonder about signal stationarity. Next either a non linear (thus like a power law) trend or a periodic background could be sub-

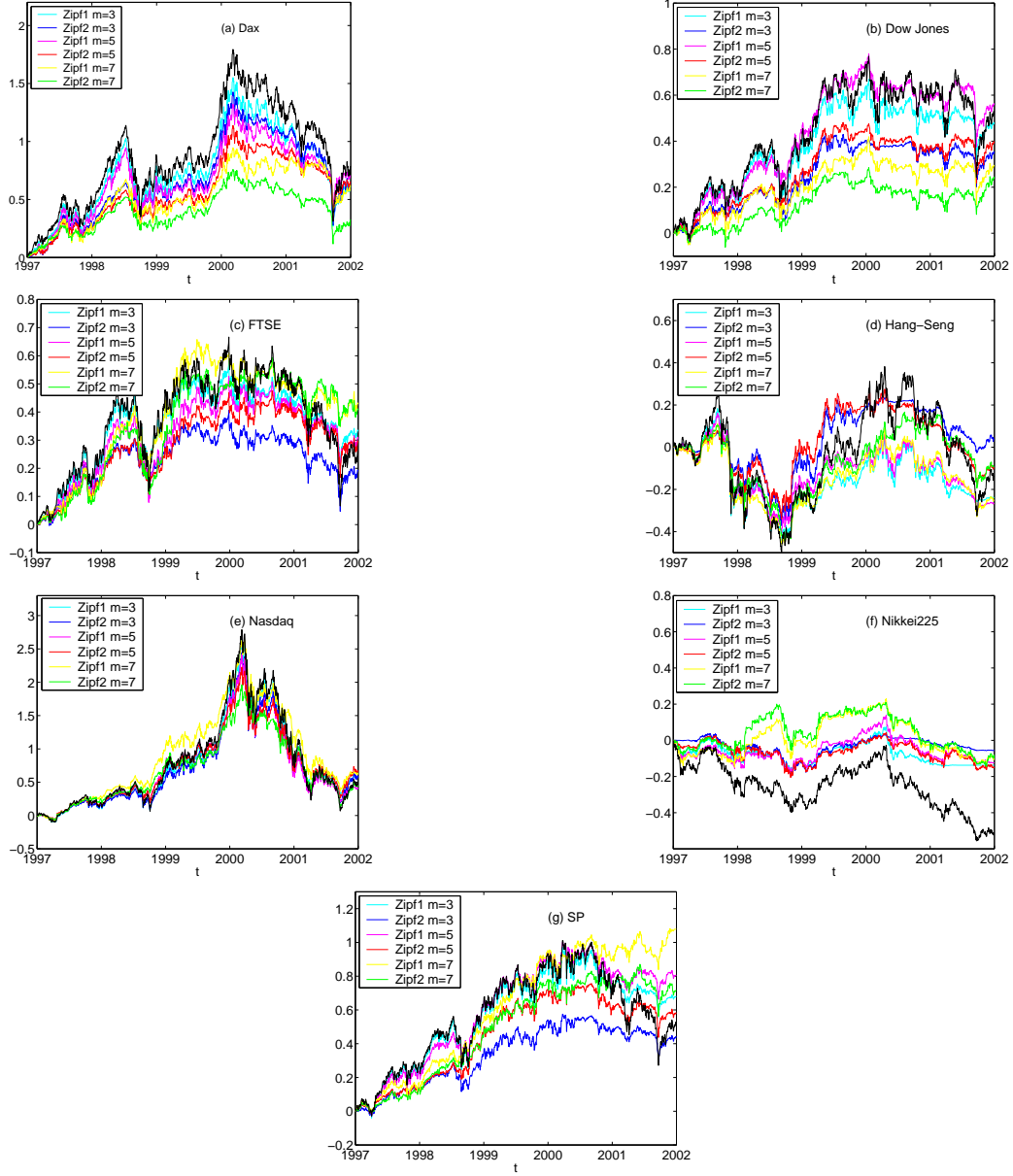


Figure 2: The time dependent returns for $Z1$ and $Z2$ in the case $m=3$, 5, and 7, and for $k=2$ for the seven considered financial indices

tracted from the raw signal, and the Zipf exponent time variation examined. Many other strategies are also available.

In summary, we have translated seven financial index signals each into a text based of two letters u and d , according to the fluctuations as in a corresponding random walk. We have calculated the Zipf exponent(s) giving the relationship between the frequency of occurrence of words of length $m < 8$ made of such "letters" for a binary alphabet. We have introduced considerations based on the notion of a local (or "time" dependent) Zipf law (and exponent). We have imagined two simple investment strategies taking into account the linear trend of the biased signal or not, and have reported the time dependence of the returns.

Acknowledgments

We thank K. Ivanova for stimulating discussions and comments.

References

- [1] A.G. Ellinger, *The Art of Investment*, (Bowers & Bowers, London, 1971); E.F. Fama, J. Finance **45**, 1089 (1990)
- [2] M. Ausloos and K. Ivanova, "Mechanistic approach to generalized technical analysis of share prices and stock market indices", *Eur. Phys. J. B* **27**, 177 (2002)
- [3] G.K. Zipf, *Human Behavior and the Principle of Least Effort*, (Addison-Wesley, Cambridge, MA, 1949)
- [4] B.J. West and W. Deering, *Phys. Rep.* **246**, 1 (1994)
- [5] N. Vandewalle and M. Ausloos, "Coherent and random sequences in financial fluctuations", *Physica A* **246**, 454 (1997)
- [6] K. Ivanova and M. Ausloos, "False euro (FEUR) exchange rate correlated behaviors and investment strategy", *Eur. Phys. J. B* **20**, 537 (2001)

- [7] see [http : //linkage.rockefeller.edu/wli/zipf/](http://linkage.rockefeller.edu/wli/zipf/) and in particular : B. Hill, "The Rank-Frequency Form of Zipf's Law", *J. Am. Stat. Assoc.* **69**, 1017 (1974); B. Hill and M. Woodroffe, "Stronger Forms of Zipf's Law", *J. Am. Stat. Assoc.* **70**, 212 (1975); M. A. Montemurro, "Beyond the Zipf-Mandelbrot law in quantitative linguistics", *Physica A* **300**, 567 (2001); A.N. Pavlov, W. Ebeling, L. Molgedey, A.R. Ziganshin and V.S. Anishchenko, "Scaling features of texts, images and time series", *Physica A* **300**, 310-324 (2001); B. Gutenberg and R.F. Richter, "Frequency of earthquakes in California", *Bull. Seis. Soc. Amer.* **34**, 185 (1944); D. Sornette, L. Knopoff, Y.Y. Kagan, and C. Vanneste, "Rank-ordering statistics of extreme events: application to the distribution of large earthquakes", *J. Geophys. Res.* **101**, 13883 (1996); Y.G. Ma, "Zipf's law in the liquid gas phase transition of nuclei", *Eur. Phys. J. A* **6**, 367 (1999); J.R. Piqueira, L.H. Monteiro, T.M. de Magalhaes, R.T. Ramos, R.B. Sassi, and E.G. Cruz, "Zipf's law organizes a psychiatric ward", *J. Theor. Biol.* **198**, 439 (1999); W. Li, "Zipf's law in importance of genes for cancer classification using microarray data", arxiv.org e-print, physics/0104028; J. Kalda, M. Sakki, M. Vainu, and M. Laan, "Zipf's law in human heartbeat dynamics", arxiv.org e-print , physics/0110075; X. Gabaix, "Zipf's Law For Cities: An Explanation", *Quart. J. Econ.* **114**, 739 (1999); X. Gabaix, "Zipf's Law and the Growth of Cities", *American Economic Review Papers and Proceedings* **89**, 129 (1999)
- [8] B. Mandelbrot, "The Pareto-Levy Law and the Distribution of Income", *Intern. Econ. Rev.* **1**, 79 (1960); S. Solomon and P. Richmond (2000), "Stability of Pareto-Zipf law in non-stationary economics", arxiv.org e-print , cond-mat/0012479.
- [9] M.H.R. Stanley, S.V.Buldyrev, S.Havlin, R.Mantegna, M.A.Salinger, and H.E.Stanley, "Zipf plots and the size distribution of Firms", *Eco. Lett.* **49**, 453 (1995); J.J. Ramsden and Gy. Kiss-Haypal, "Company size distribution in different countries", *Physica A* **277**, 220 (2000)
- [10] M. Marsili and Y.-C. Zhang, "Interacting Individuals Leading to Zipf's Law", *Phys. Rev. Lett.* **80**, 2741 (1998)
- [11] N. Vandewalle and M. Ausloos, "The n-Zipf analysis of financial data series and biased data series", *Physica A* **268**, 240 (1999)

- [12] K. Okuyama, M. Takayasu, and H. Takayasu, "Zipf's Law in income distribution of companies", *Physica A* **269**, 125 (1999)
- [13] H Aoyama, W Souma, Y Nagahara, M P Okazaki, H Takayasu, M Takayasu, "Pareto's law for income of individuals and debt of bankrupt companies", *Fractals* **8**, 293 (2000)
- [14] M. Ausloos and K. Ivanova, "Precise (m,k)-Zipf diagram analysis of mathematical and financial time series when $m = 6$, $k = 2$ ", *Physica A* **270**, 526 (1999)
- [15] M. Ausloos, "Statistical Physics in Foreign Exchange Currency and Stock Markets", A. Pękalski Ed., Proc. Ładź Zdrój Conference on "Exotic Statistical Physics", *Physica A* **285**, 48 (2000)
- [16] Ph. Bronlet and M. Ausloos, "Generalized (m,k)-Zipf law for fractional Brownian motion-like time series with or without effect of an additional linear trend", *Int. J. Mod. Phys. C* **14**, xx (2003)
- [17] H.E. Hurst, *Trans. Amer. Soc. Civil Eng.* **116**, 770 (1951).
- [18] P.S. Addison, *Fractals and Chaos*, (Institute of Physics, Bristol, 1997)
- [19] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, "Testing for nonlinearity in time series: The method of surrogate data", *Physica A* **58**, 77 (1992)

Figure Captions

Figure 1 – The (DAX, FTSE; DJIA, NASDAQ, S&P500; Hang-Seng and Nikkei 225) indices have been obtained from [http : //finance.yahoo.com/](http://finance.yahoo.com/). They contain *ca.*1250 data points, from Jan. 01, 1997 till Dec. 31, 2001

Figure 2 – The time dependent returns for $Z1$ and $Z2$ in the case $m=3$, 5, and 7, and for $k=2$ for the seven considered financial indices